美团四层负载均衡 - MGW

王伟@云计算部 20161008

自我介绍

2013/8月 ~ 2015/4月

深信服电子科技 (应用交付部)

1. Linux kernel开发（网络部分）
2. SSL/TLS 开发（openssl）

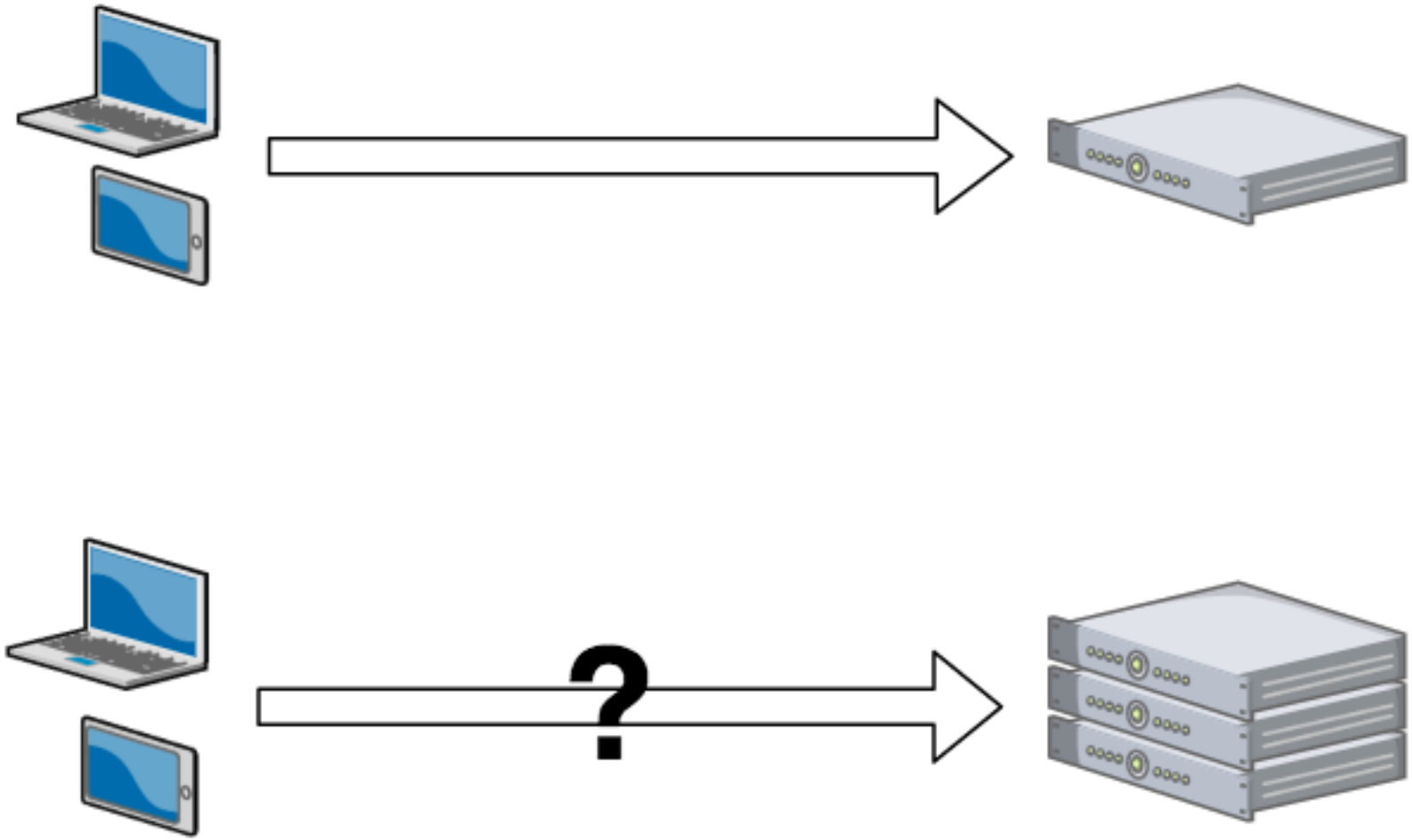
2015/4月 ~ 至今

美团（云计算部）

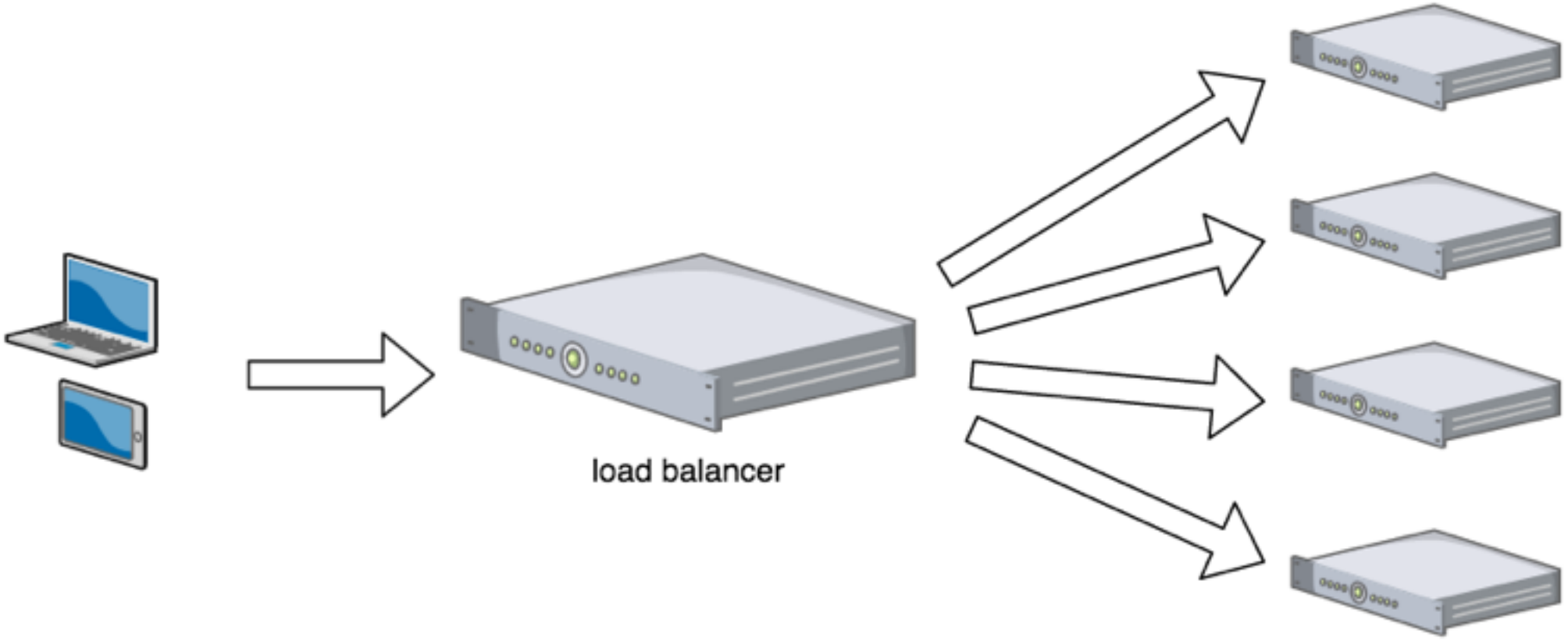
1. 四层负载均衡网关（MGW)
2. Linux kernel开发

目录

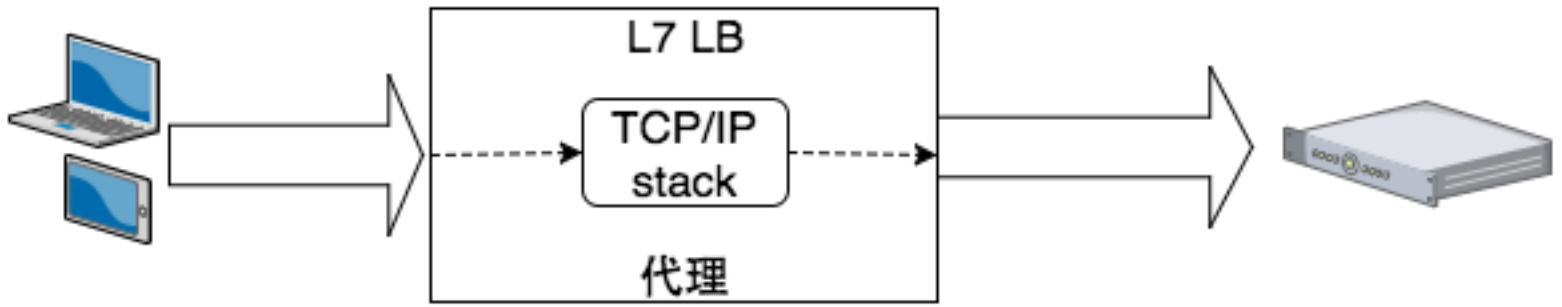
什么是负载均衡?



什么是负载均衡?



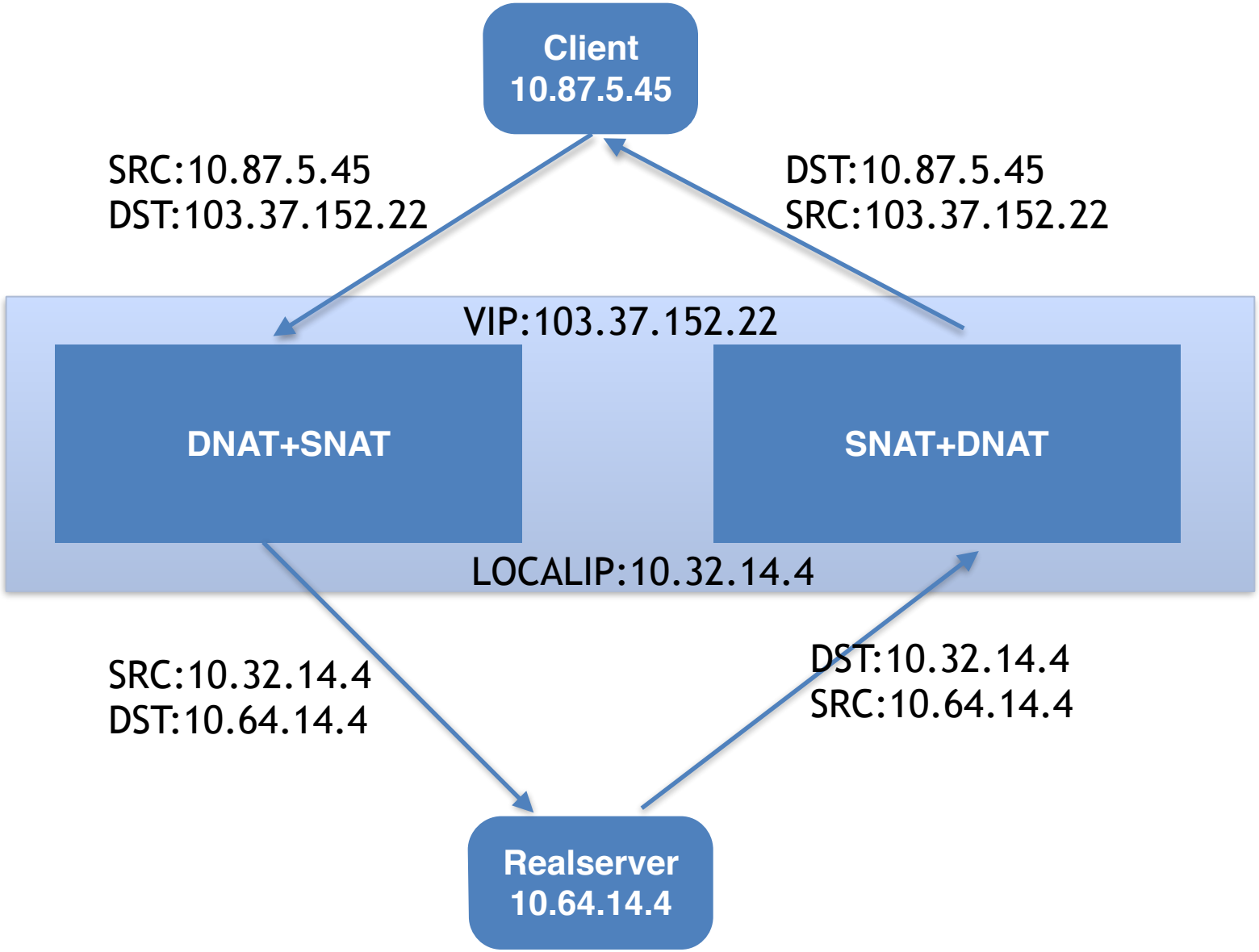
四层负载均衡 and 七层负载均衡?



转发模式对比

模式	优点	缺点
DR(三角传输）	1. 应用直接将应答发给客户端，性能好	1. 必须在一个二层 2. 应用服务器需要配置VIP
NAT（DNAT）	1. 应用服务器无需做配置	1. 负载均衡必须以网关形式存在
TUNNEL	1. 和DR一样，应用直接将应答发给客户端，性能好。	1. 对应用服务器要求高，需要支持tunnel 2. 应用服务器需要配置vip
FULLNAT（SNAT+DNAT)	1. 应用服务器无需做配置 2. 对网络环境要求比较低	1. 丢失client ip

转发模式 - FULLNAT



目录

负载均衡

硬件负载均衡	软件负载均衡
	 LVS
CITRIX	NGINX
	HAPROXY

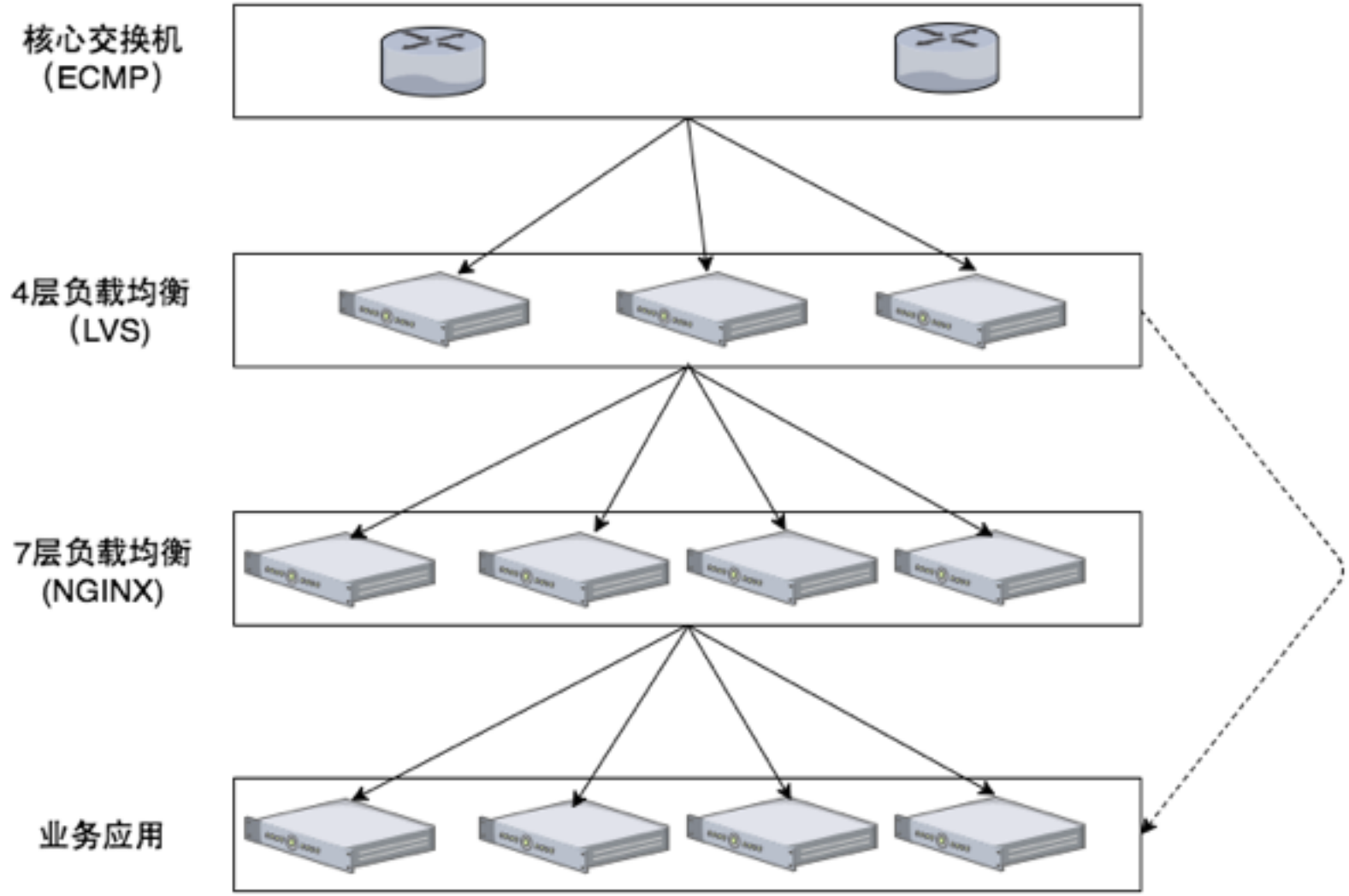
硬件负载均衡

1. 硬件成本

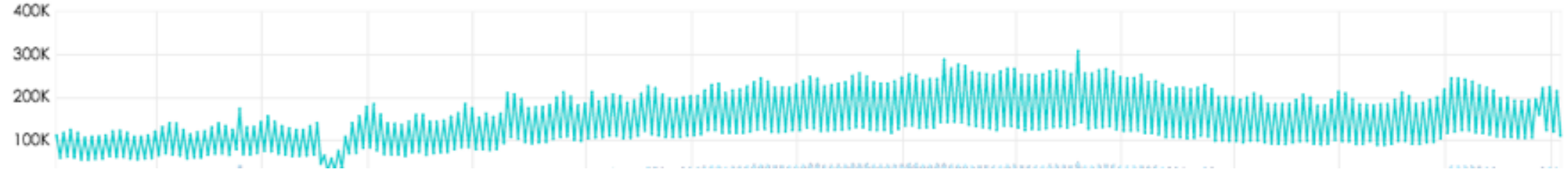
2. 人力成本

3. 时间成本

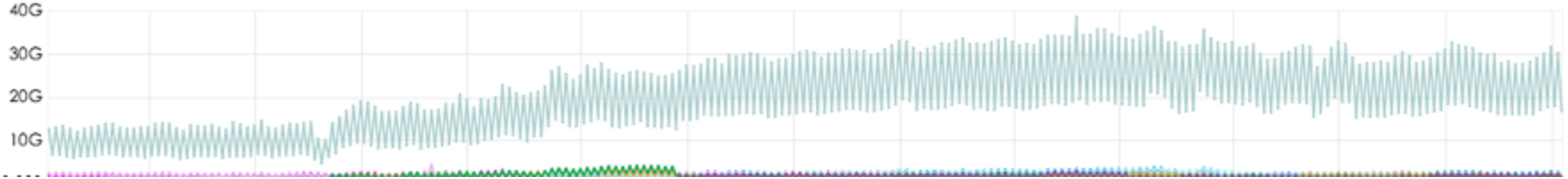
美团早期负载均衡结构



美团流量增长情况



新建连接数（cps）100w -> 300w



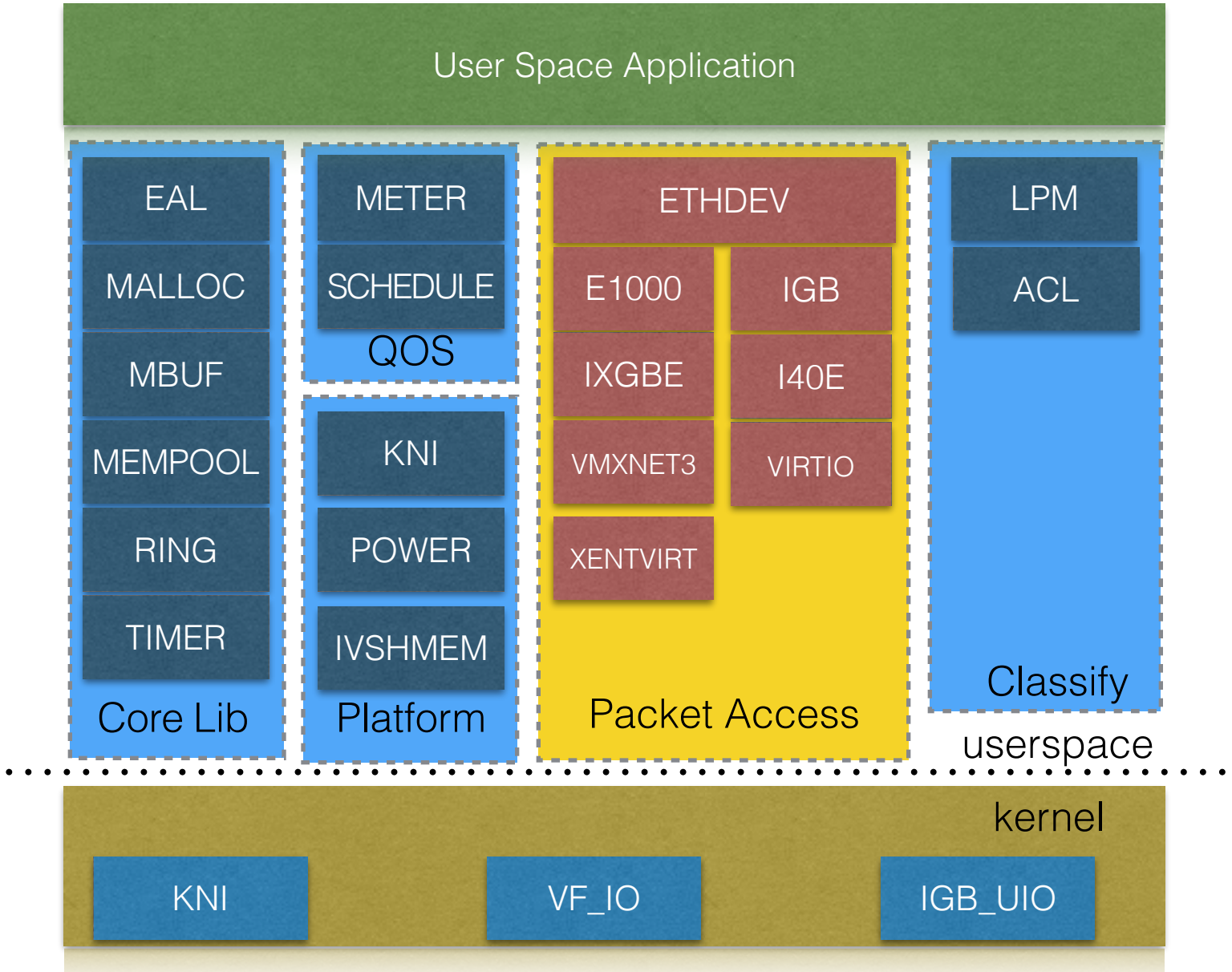
吞吐量（bps）13G -> 38G

❗3倍的流量增长，LVS性能不足，故障率增加

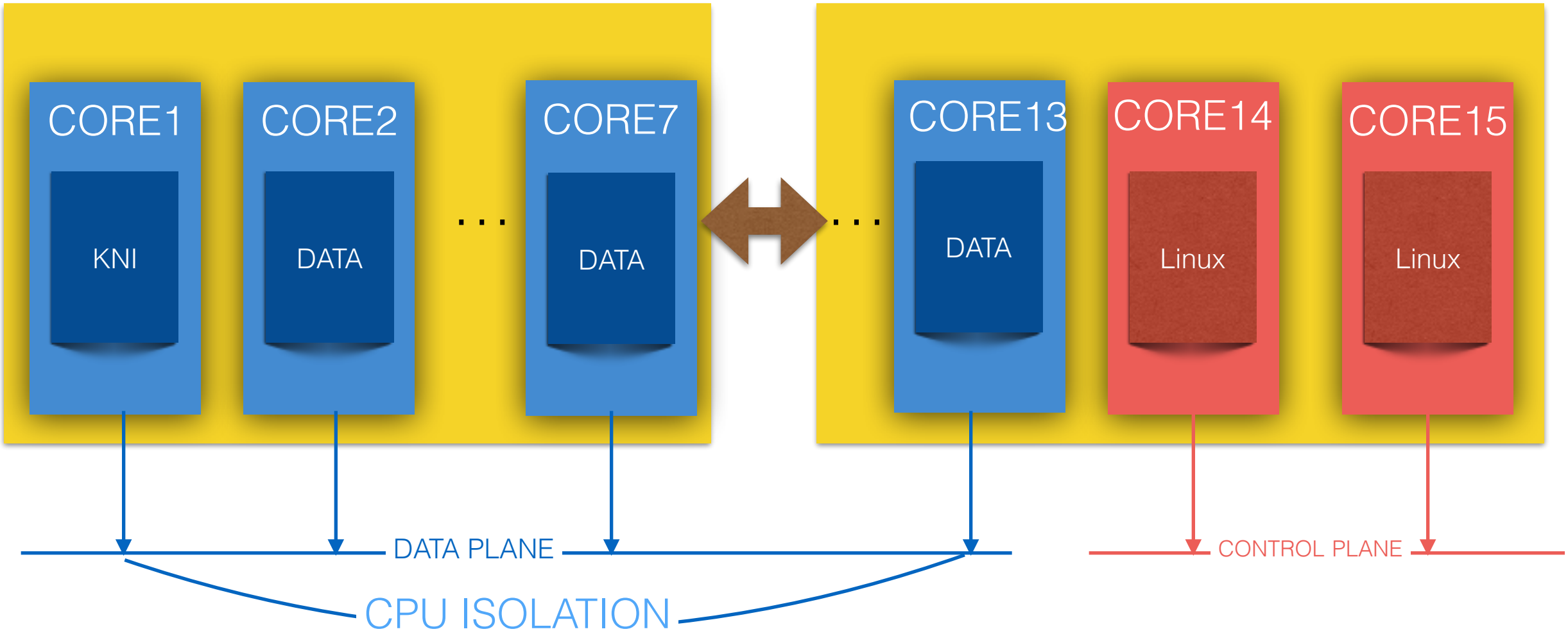
问题所在

1. 中断 → 1. PMD驱动
2. 过长的协议栈路径 → 2. kernel bypass
3. 锁 → 3. 无锁的设计
4. 上下文切换 → 4. CPU绑定、隔离

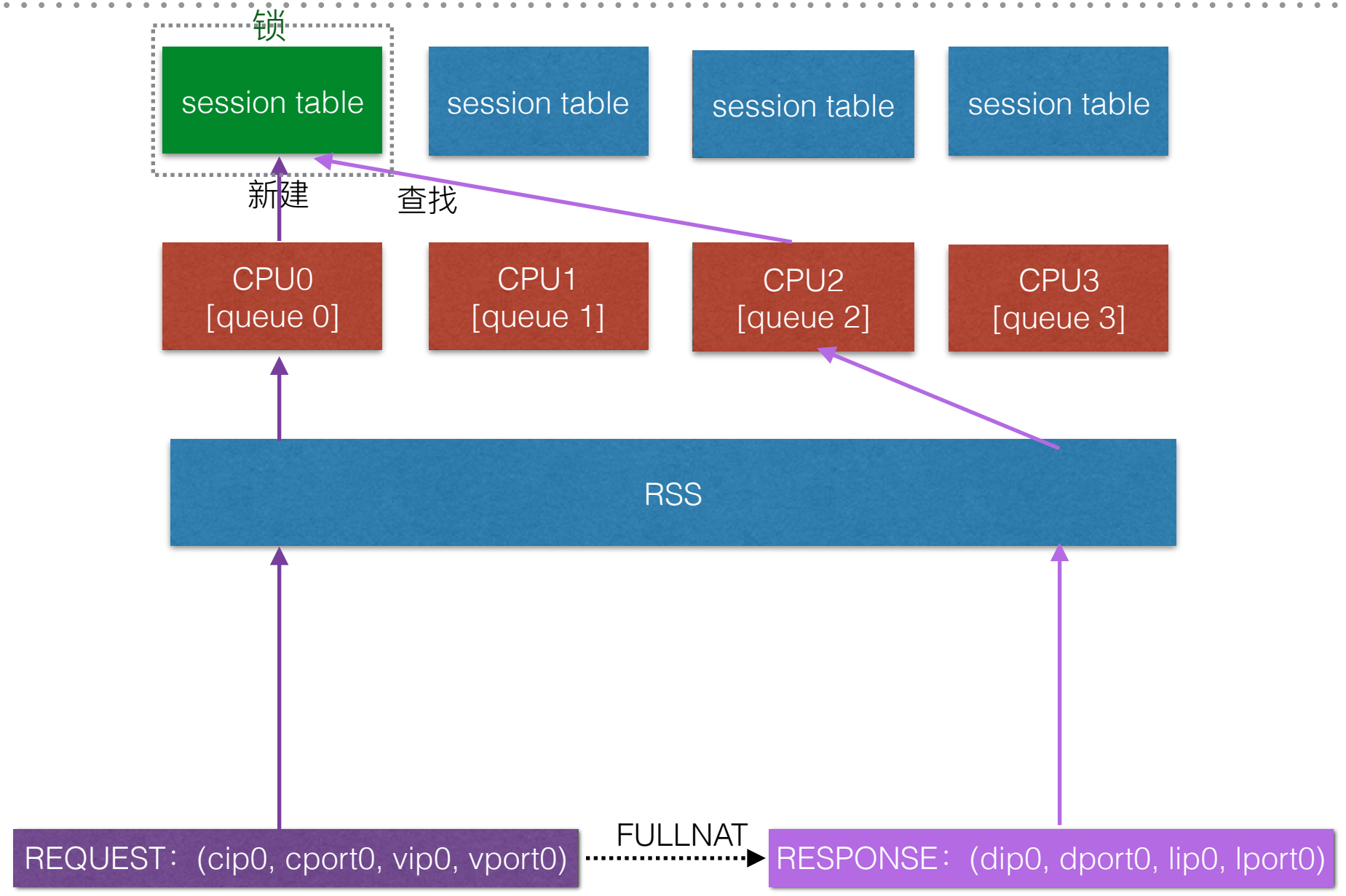
PMD驱动、kernel bypass -> DPDK



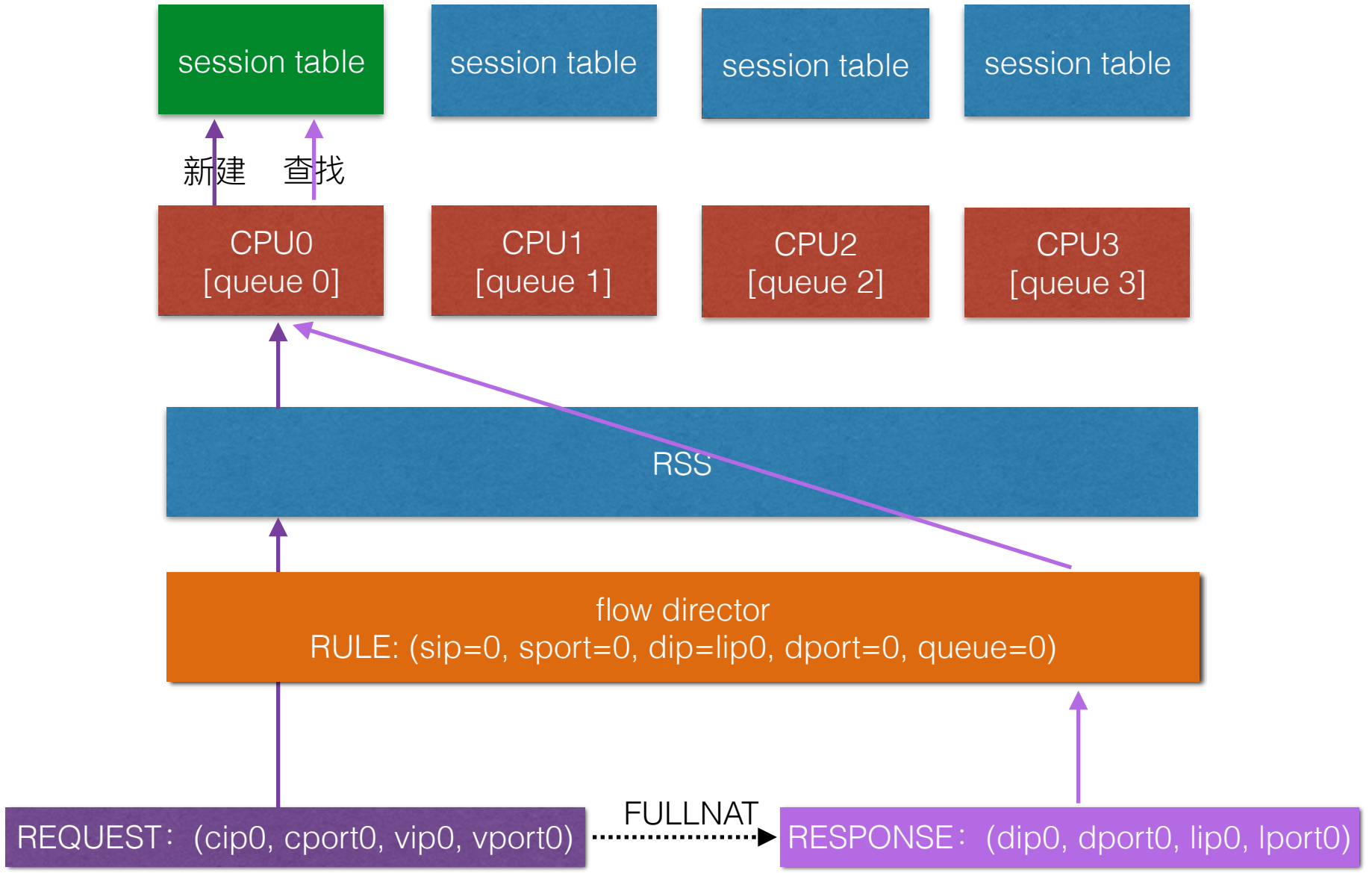
上下文切换


CORE1
KNI
CORE2
DATA
CORE7
DATA
CORE13
DATA
CORE14
Linux
CORE15
Linux
DATA PLANE
CONTROL PLANE
CPU ISOLATION

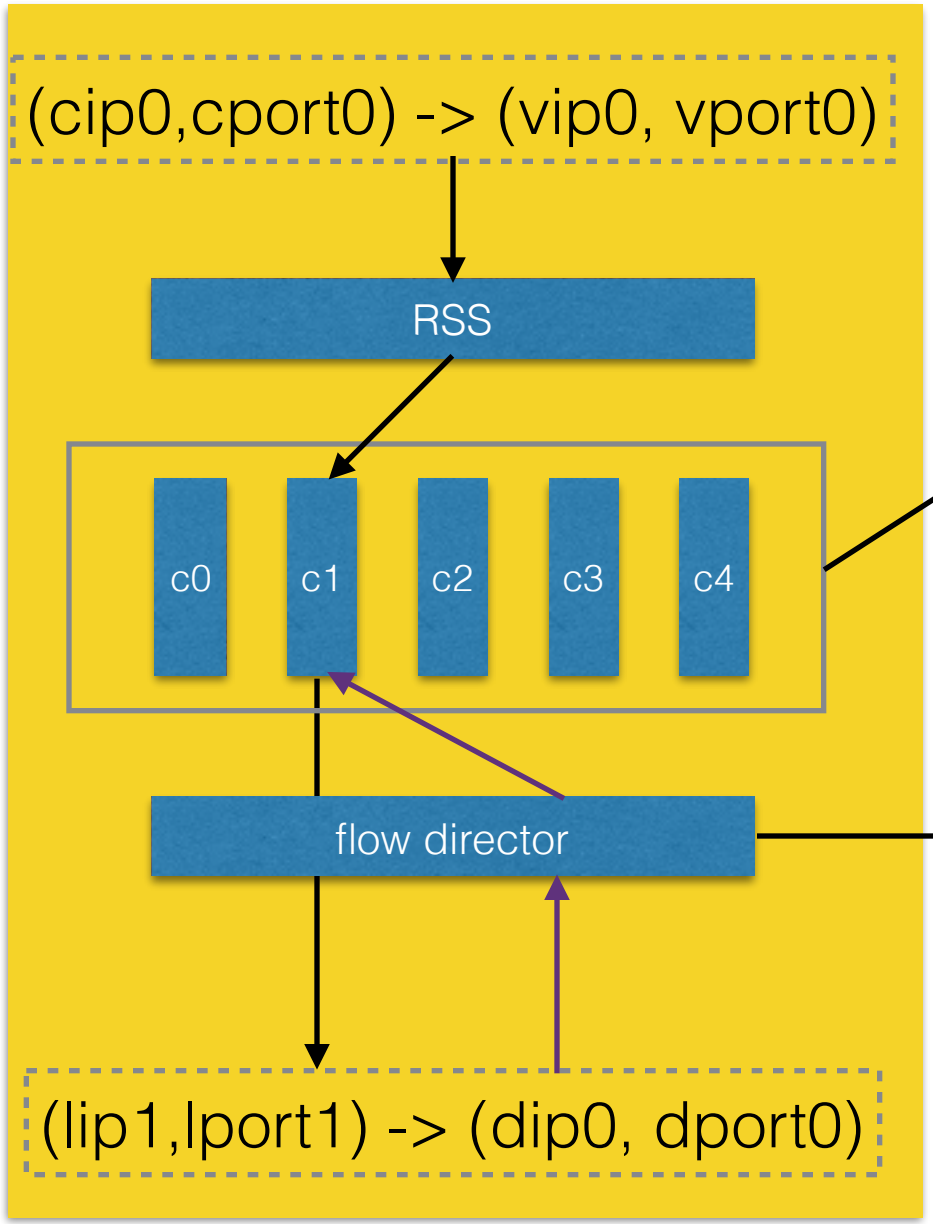
无锁的设计 - 地址转换的问题



无锁的设计 - 地址转换的问题



无锁的设计



lip（localip）池

CPU0	CPU1	CPU2	CPU3
lip0	lip1	lip2	lip3
lip4	lip5	lip6	lip7
lip8	lip9	lip10	lip11
lip12	lip13	lip14	lip15

flow director 规则

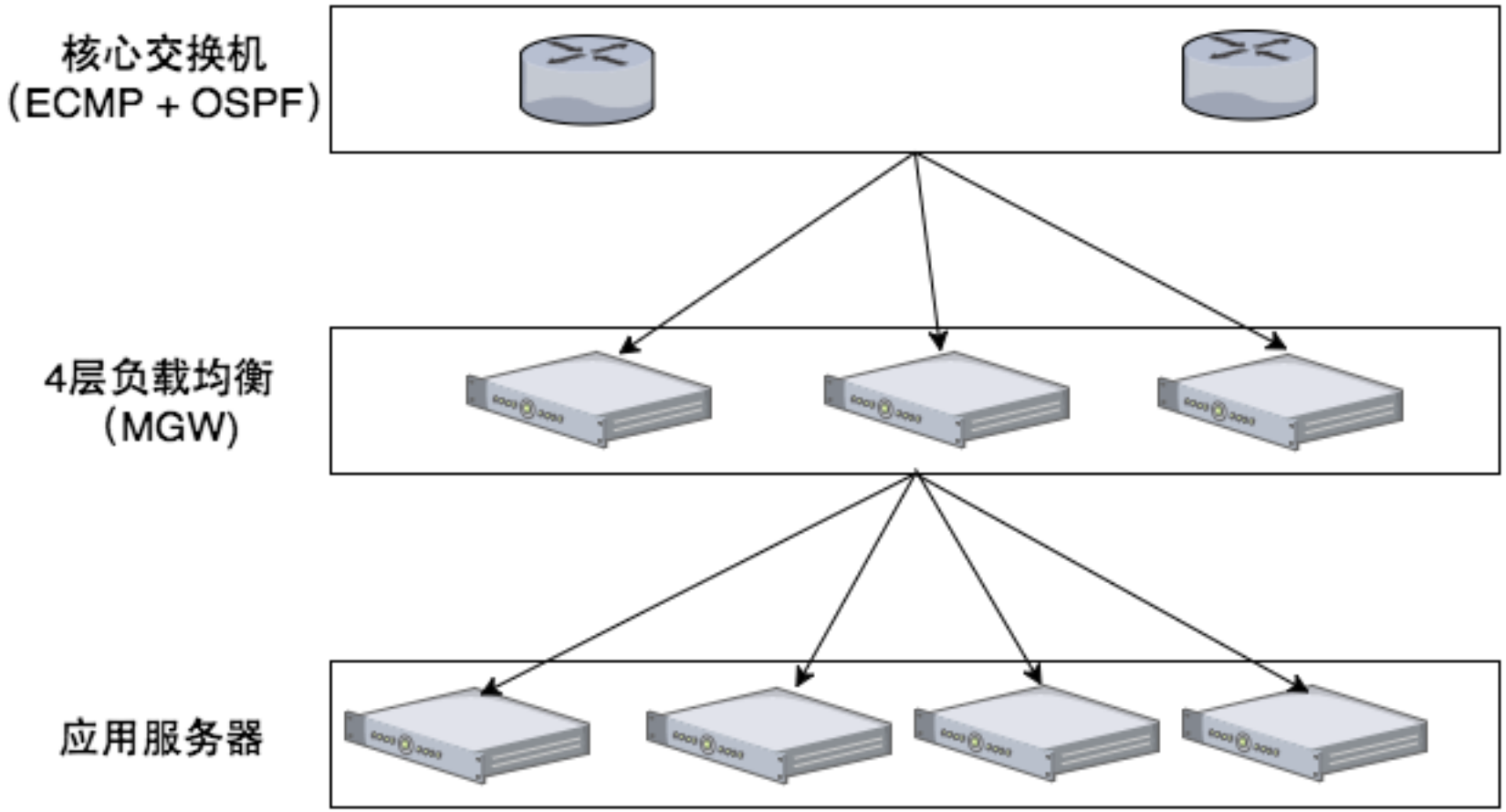
规则
(sip=0, sport=0, dip=lip0, dport=0, queue=0)
(sip=0, sport=0, dip=lip1, dport=0, queue=1)
(sip=0, sport=0, dip=lip6, dport=0, queue=2)
(sip=0, sport=0, dip=lip7, dport=0, queue=3)

性能测试

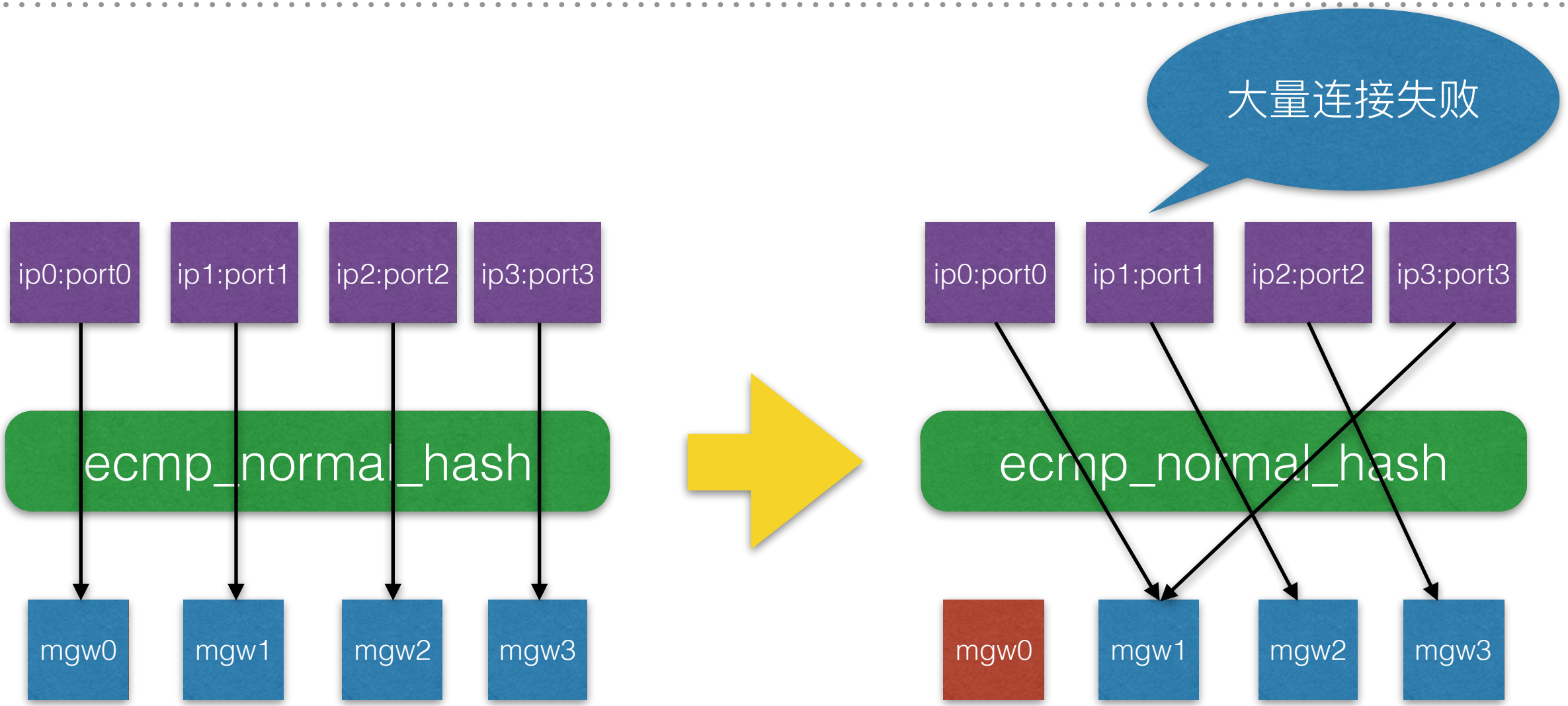
测试项	性能参数
SYNPROXY	2800w pps (67%CPU)
http pps	in 850w out 960w
http bps (64bytes response)	in 6G out 10G
http qps	380w

目录

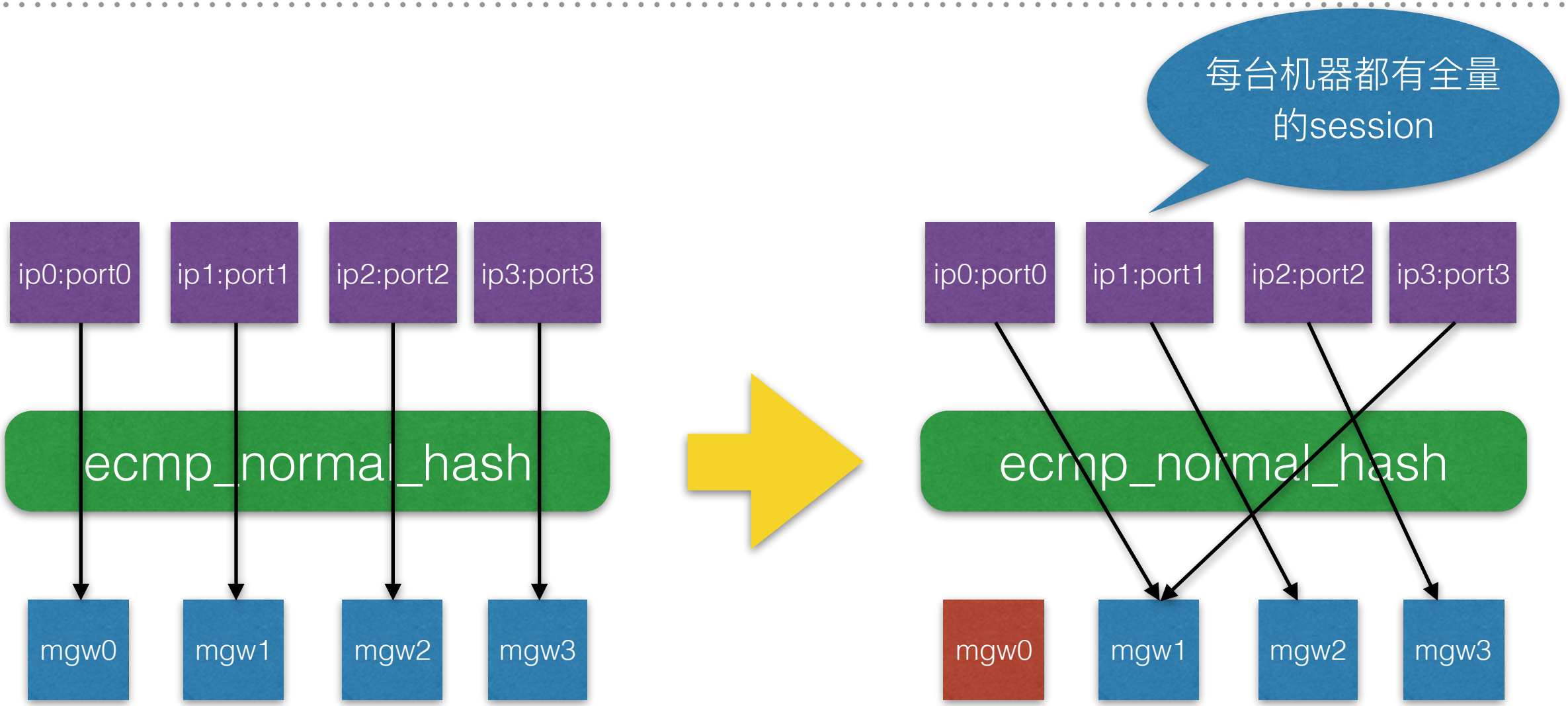
高可靠



机器下线导致的问题


大量连接失败
ip0:port0
ip1:port1
ip2:port2
ip3:port3
ecmp_normal_hash
mgw0
mgw1
mgw2
mgw3
ip0:port0
ip1:port1
ip2:port2
ip3:port3
ecmp_normal_hash
mgw0
mgw1
mgw2
mgw3

session同步

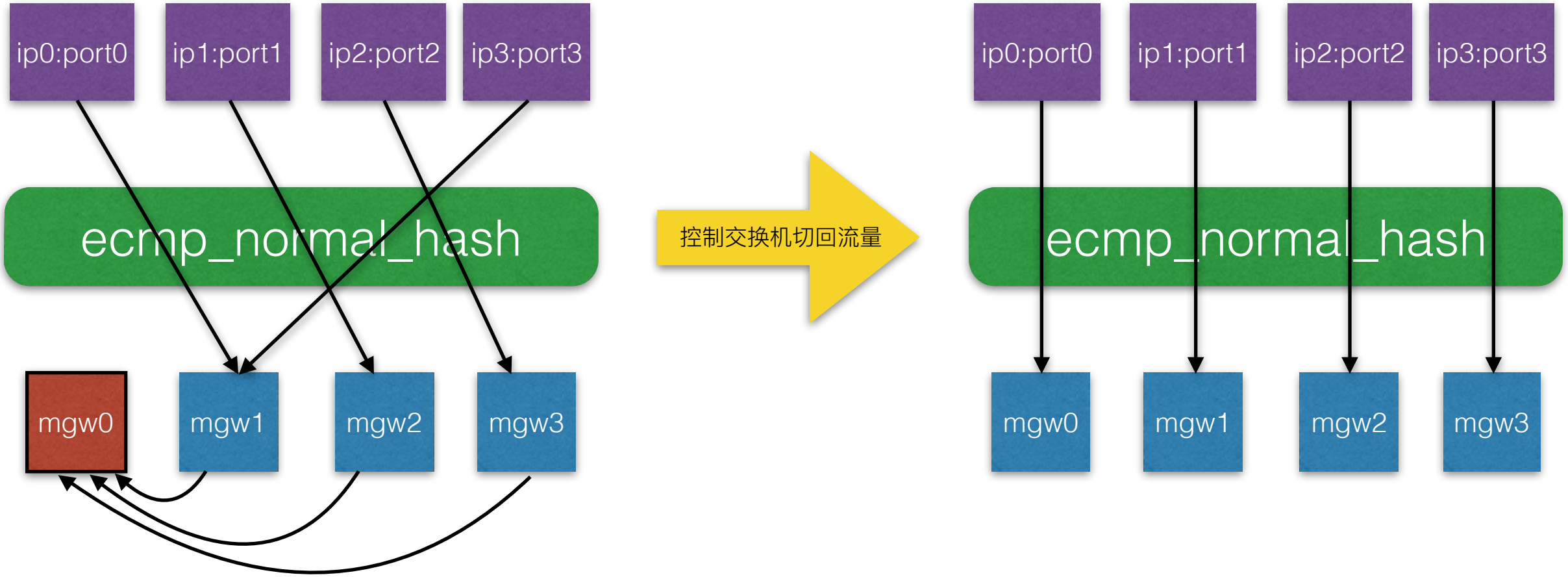

每台机器都有全量的session
ip0:port0
ip1:port1
ip2:port2
ip3:port3
ecmp_normal_hash
mgw0
mgw1
mgw2
mgw3
ip0:port0
ip1:port1
ip2:port2
ip3:port3
ecmp_normal_hash
mgw0
mgw1
mgw2
mgw3

故障检测与故障切换

故障切换
交换机侧不使用虚拟接口
半秒一次机器健康自检
检测到故障自动给网口断电
捕获异常信号，物理网口断电

故障切换效果	
测试程序发包间隔	100ms
升级操作丢包	0
主程序故障丢包	0
其他异常（网线等）	500ms

故障恢复与扩容

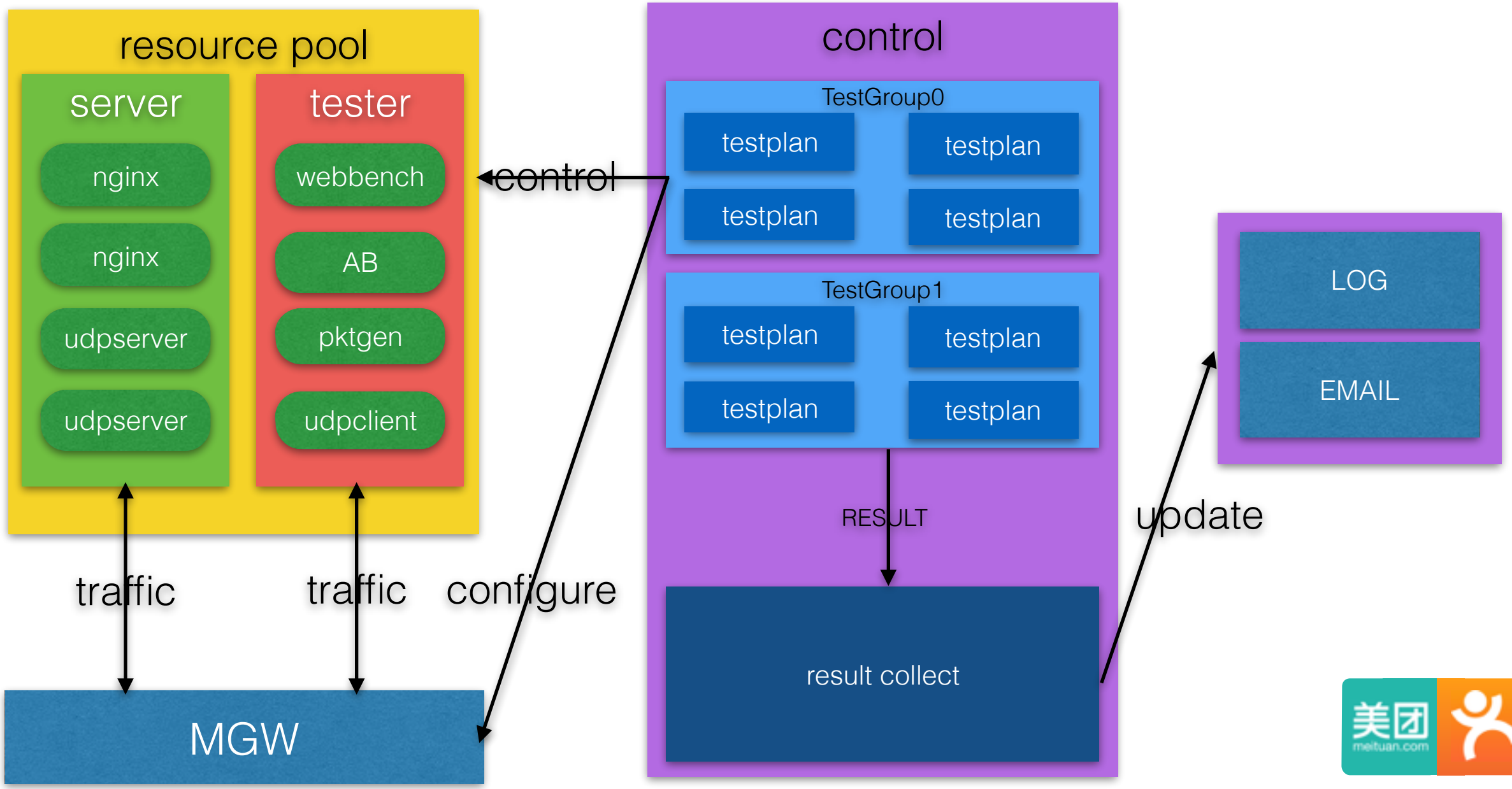


批量session同步同步旧session，同时通过增量同步接收新session

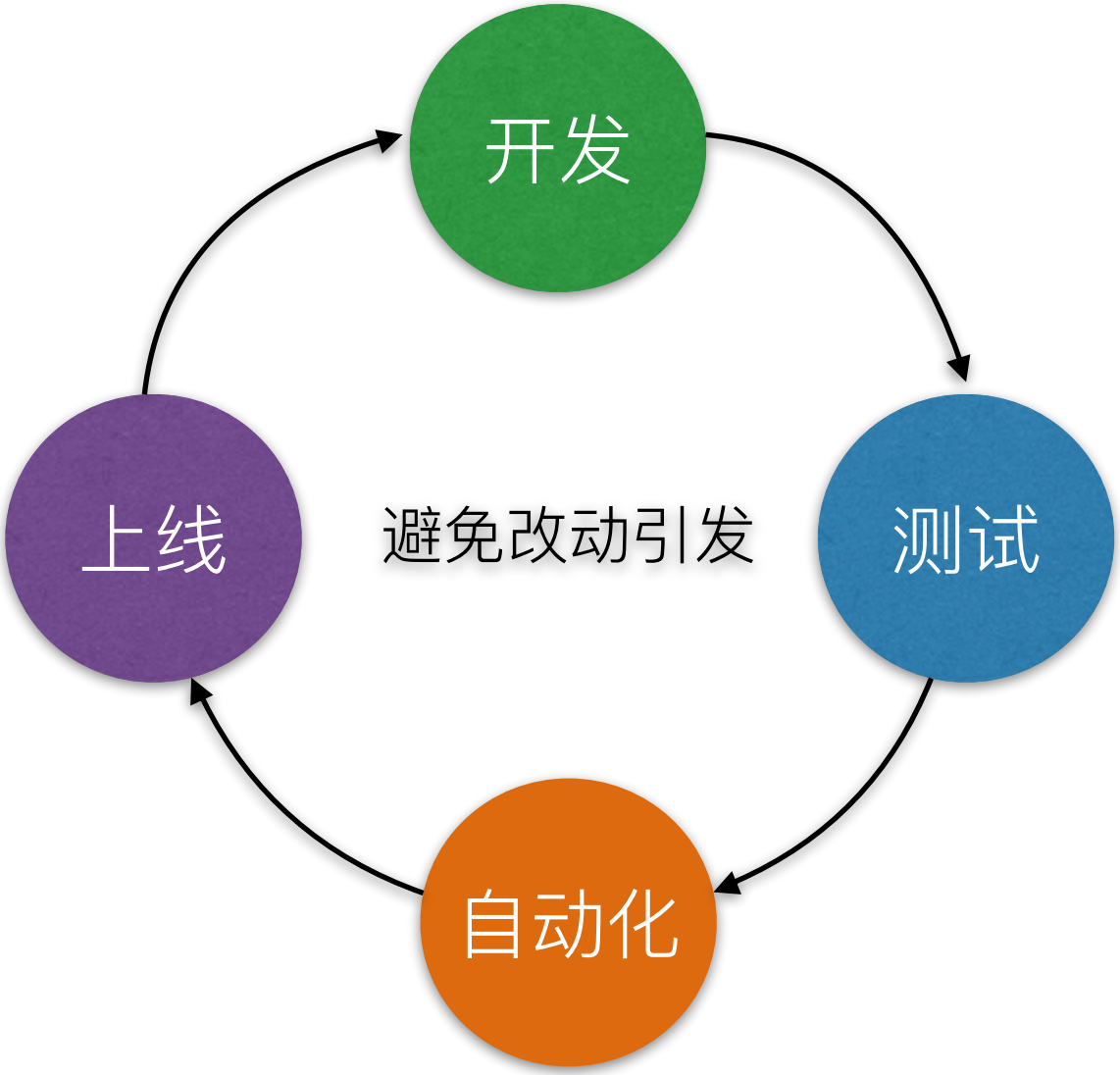
MGW单机可靠性

自动化测试

自动化测试平台


resource pool
server
nginx
nginx
udpserver
udpserver
tester
webbench
AB
pktgen
udpclient
control
control
TestGroup0
testplan
testplan
testplan
testplan
TestGroup1
testplan
testplan
testplan
testplan
RESULT
result collect
LOG
EMAIL
update
traffic
traffic
configure
MGW

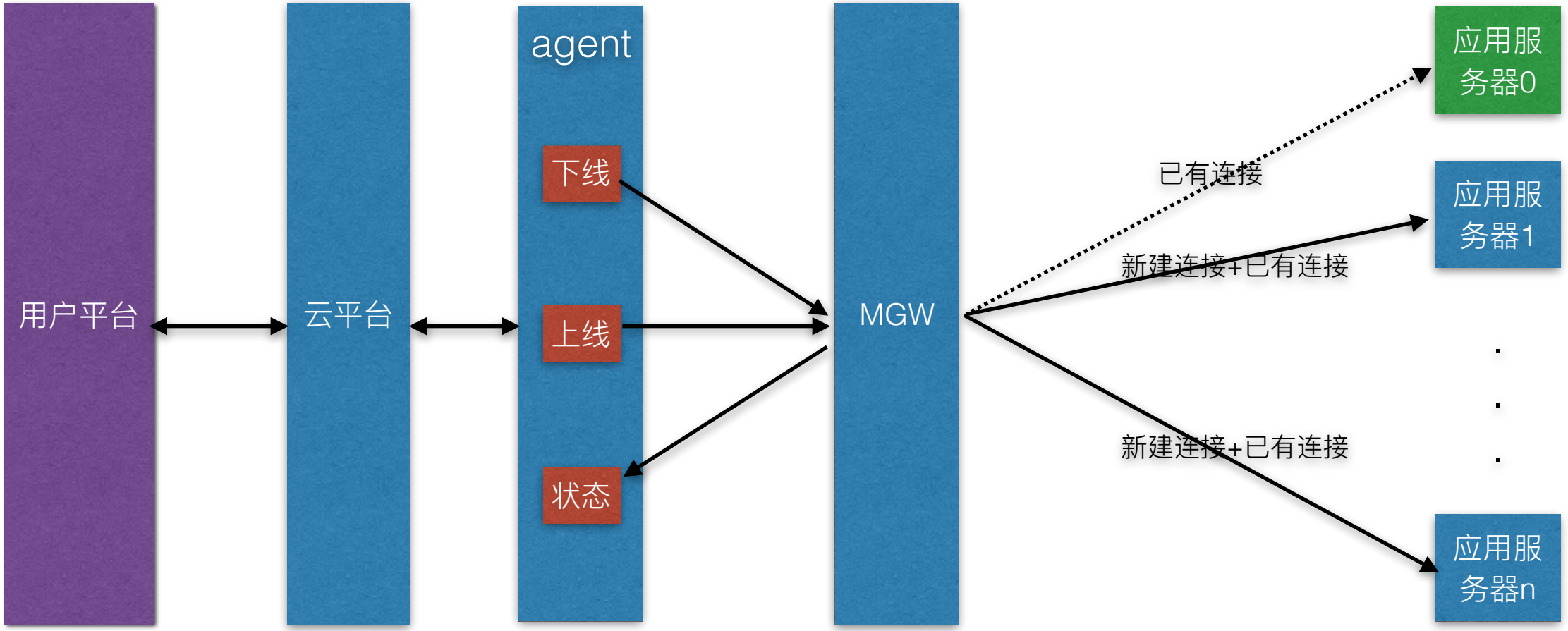
自动化测试平台



应用服务器可靠性

节点平滑下线

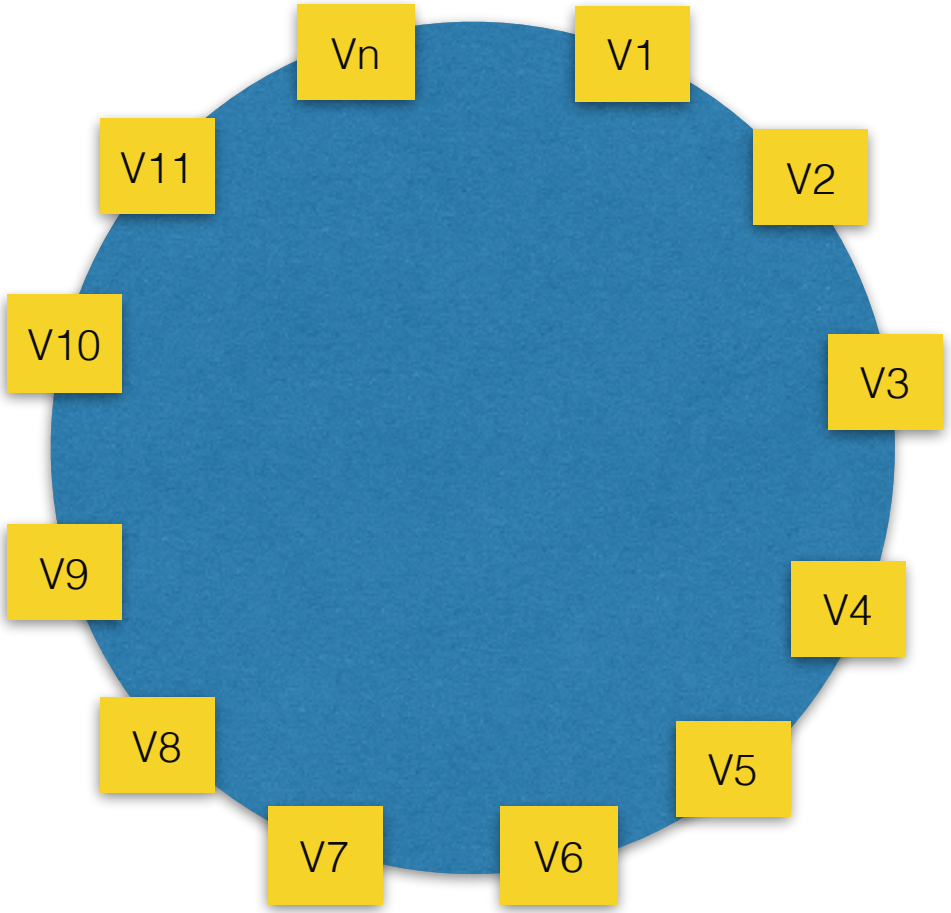
节点平滑下线

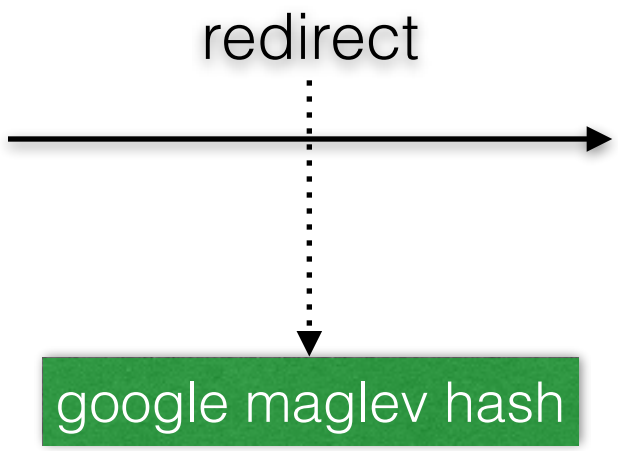

用户平台
云平台
agent
下线
上线
状态
MGW
已有连接
新建连接+已有连接
新建连接+已有连接
应用服务器0
应用服务器1
应用服务器n

应用服务器可靠性

一致性源ip hash调度器

一致性hash调度器

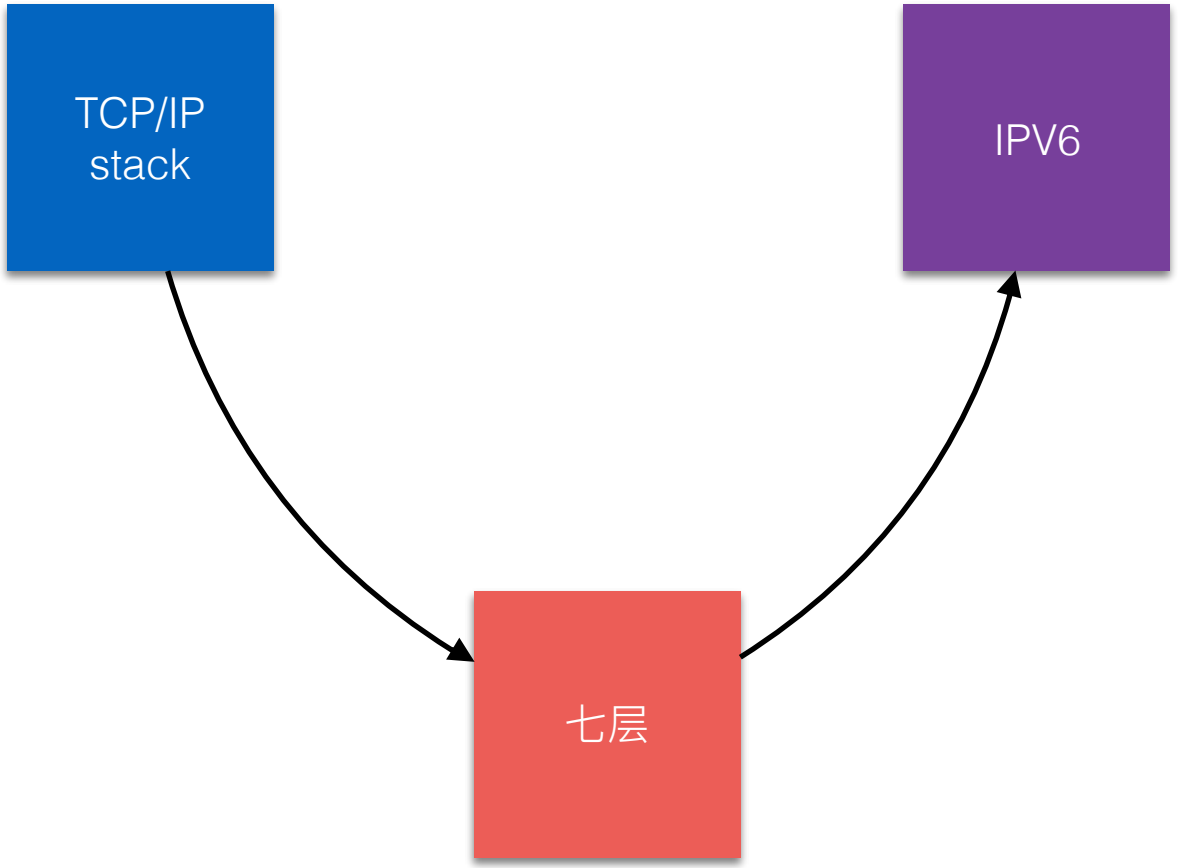




V1	node0
V2	node1
V3	node2
V4	node3
V5	node0
V6	node1
V7	node2
V8	node3
V9	node0

目录

技术展望



谢谢大家

Q&A